FAMILIAR SPEAKER RECOGNITION

Stanley J. Wenndt

Air Force Research Laboratory
Rome, NY 13440, USA
Stanley.Wenndt@rl.af.mil

Ronald L. Mitchell

Clarkson University
Potsdam, NY 13699, USA
mitcherl@clarkson.edu

ABSTRACT

Speaker recognition by machines can be quite good for large groups as seen in NIST speaker recognition evaluations. However, speaker recognition by machine can be fragile for changing environments. This research examines how robust humans are for recognizing familiar speakers in changing environments. Additionally, bandlimited noise was used to try to learn what frequency regions are important for human listeners to recognize familiar speakers.

Index Terms— Speaker Familiarity, Voice Recognitions by Humans

1. INTRODUCTION

While speaker recognition by machines can be quite good for large groups, cross-conditions between the training and testing data can still cause a drop in recognition performance. Additionally, every speaker model is built the same way with the same features and same steps being used to train every speaker model. This research takes a cursory look at how robust humans are at identifying familiar speakers in changing environments.

For speaker identification, there may be many clues that convey the speaker's identity. Figure 1 shows a possible approach that humans may use for speaker identification. It was derived by analyzing potential linguistic and non-linguistic cues. It is not expected that all the information is needed or used by humans, but instead, a systematic approach is used. If the person is very familiar to the listener, then the word choice or non-linguistic information (such as a laughter or stutter) may allow the listener to quickly identify the speaker. For speakers that are less familiar to the listener, then a multi-step process may be needed such as narrowing the possible choices by the gender, age, accent, etc., until enough information is gathered to make a successful identification.

In [1], the research showed that female speakers were recognized better by both male and females listeners. Thus, the cues needed for female recognition may be different than those needed for male identification. Or, the cues for males may be less distinct. For homogenous groups, it is expected that more cues would be required to make an identification. For machine algorithms, it's hard to correlate which cues may be captured by the cepstral feature.

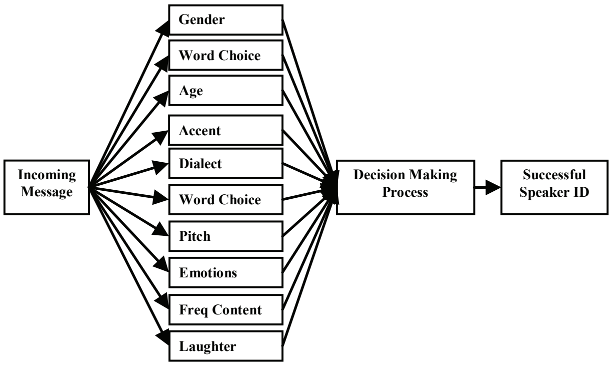

Figure 1: Potential Cues for Speaker Recognition by Humans.

Part of the motivation for this work began by looking at speech from a psychoacoustic perspective, that is, what information affects the recognition of familiar speakers. In [2], speech stimuli (sentences) were filtered into five adjacent bands of equal intelligibility where each band provided equal contribution for the overall intelligibility of the stimuli in a quiet environment. The low band of 111-561 Hz is only 596 Hz wide while the high band of 2807-11000 Hz is 7439 Hz wide; yet their overall contribution to intelligibility is equal. The additive noise was matched to the long-term average speech spectrum (LTASS) [3] of the stimuli. This is an important step to ensure that the additive noise for each band has the same signal-to-noise ratio (SNR). The research found that listeners placed greater importance on the second and fifth band region in degraded noise environments for sentences. If different region carry different information and importance for speech intelligibility, then, likewise one might expect different regions to carry different information and importance for speaker recognition.

Speech can be modeled as a source-filter type problem and both the source and filter can provide cues for speaker recognition. In [4], the author eliminated the inter-speaker variability of the source by using an electro-larynx. There were 10 male speakers and 10 female speakers. The speakers were instructed how to produce non-phonated speech using an electro-larynx. For both male and female speakers, the fundamental frequency of the electro-larynx was 85 Hz with a jitter of ±3Hz. The experimental task was to listen to two voices and decide if the voices were the same or not. Results for over 1100 speaker comparisons yielded a success rate of greater than 90%. Thus, there are still sufficient cues, aside from the glottal excitation, to allow for speaker discrimination. Interesting, there were significantly less errors for

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAY 2012**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Familiar Speaker Recognition**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory Rome, NY 13440, USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

See also ADA561051. AOARD-CSP-111007 International Conference on Acoustics, Speech and Signal Processing (37th) (ICASSP 2012) Held in Kyoto, Japan on March 25-30, 2012. U.S. Government or Federal Purpose Rights License., The original document contains color images.

14. ABSTRACT

Speaker recognition by machines can be quite good for large groups as seen in NIST speaker recognition evaluations. However, speaker recognition by machine can be fragile for changing environments. This research examines how robust humans are for recognizing familiar speakers in changing environments. Additionally, bandlimited noise was used to try to learn what frequency regions are important for human listeners to recognize familiar speakers.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**SAR**	**4**	

male-male pairs compared to female-female pairs. Once again, it seems that different speakers or different groups of speakers have different cues towards their identity.

Speaker recognition by listeners has been studied for a while [5], [6]. However, the challenge with speaker recognition experiments is that it is difficult to get a large group of people who are familiar with each other. Thus, experiments tend to stay at about 10-15 speakers and similarly for the number of listeners.

The article [7] showed that the discrimination of unfamiliar speakers (speaker verification) compared to the recognition of familiar speakers (speaker identification) were distinct processes that occurred in different parts of the brain. This research is focused on the familiar speaker identification task by human listeners. Section 2 describes the audio data that is used for the listening experiments, Section 3 describes the experiments, and Section 4 examines the results along with some statistical analysis.

2. AUDIO DATA

The short-term goal of this research was to learn what frequency information is important for recognition of familiar speakers by masking out certain frequency information. The long-term goal is to use this information to develop more robust speaker recognition features. This paper used additive speech-shaped noise (LTASS) to degrade particular frequency regions of the speech signal [2], [3]. This way, the signal will still sound natural and the performance of listeners can be tied to the degradation of particular frequencies. If the performance decreases when a set of frequencies are masked by an interfering signal, it would indicate that frequency range was important. Other techniques for masking or bandlimiting certain frequencies are viable, but may change the naturalness of the audio and can alter the inherent periodicity in speech.

The audio data for playback was from a previous data collection, called MARP (Multi-session Audio Research Project), which ran from May 2005 until March 2008 (USAF IRB, Protocol F-WR-2003-0032-H) and has been valuable for research in speech processing. See [8] for a more detailed description of the MARP corpus. Part of the MARP corpus was short, spontaneous sentences. The recording sessions were designed to provide realistic recordings of short sentences (about 1-2 seconds in duration) that were elicited in a casual, informal style. Read speech tends to have different speaking rates, inflections, and emotions compared to audio that is spoken more naturally (i.e., conversational speech).

Table 1: Stimuli used for Listening Experiments.

Sentence	Sentence
1	Let’s go skiing today.
2	We’ll be leaving early tonight.
3	You’re going to go with them.
4	It’s time to go now.
5	We could get a drink.
6	I need some coffee now.
7	She was home too late.
8	He broke his lower leg.
9	We need to be careful.
10	He heard the movie was great.

From the MARP corpus, there were 25 voices (20 males and 5 females) that were familiar to the listeners of the experiments in this research. The sentences used for playback were short. The goal was to make the listening experiments challenging so that additional noise would cause degradations. Table 1 lists the short sentences that were used. The sentences were downsampled to 8000 Hz. This limits the frequency content of the audio data to 4000 Hz. While there is information past 4000 Hz for audio data, an 8000 Hz sampling rate mimics a lot of communication devices such as telephones.

3. EXPERIMENTAL SETUP

There were 17 listeners in this study which included 3 females and 14 males (USAF IRB, Protocol F-WR-2010-0028-H). The listening experiments consisted of several phases. It started with a pure tone listening test. The pure tone listening test is used to verify if the speakers have normal hearing or not. The frequencies tested were 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. Then, a training phase was administered to allow the listeners to become familiar with the experimental setup and tasks. Next, there was a baseline experiment with clean stimuli (no additive noise). This was followed by several experiments with various types of noise degradation.

For this research, normal hearing was defined as 25 dB above the ANSI hearing threshold. Of the 17 participants, 11 participants had normal hearing. The other 6 participants failed at least one frequency. For the pure tone test, both the left and right ear was tested separately. However, only the right ear was used for subsequent experiments. The results are reported in two groups: those with Normal Hearing (NH) and those with a Hearing Deficit (HD). For the training phase, additional data other than the sentences listed in Table 1 were used. For this session, the listeners had the option to repeat the audio and received feedback about whether they chose the correct speaker or not. As stated before, there were 25 voices for playback. When the listener would hear a voice, they would be required to choose from a drop-down list of all 25 voices.

Sessions 1-6 were a series of experiments that started with a clean (no noise) stimuli (Session 1), and then corrupted the stimuli with speech-shaped, additive noise [3] at various frequency bands. The goal is to discover which frequency bands are most important for the familiar speaker recognition task. Table 1 lists the session number, the noise location, and the noise level. By using speech-shaped noise, the roll-off of the additive noise closely matches the roll-off of speech. The -20 dB noise level was used to mask a frequency region by having the noise level be greater than the speech level. It was also designed to degrade the performance of speaker recognition. As seen in Table 1, for Sessions 1-5, new sentences are being used. This is to prevent a learning curve from hearing the same stimuli over and over. Session 6 (no noise) is identical to Session 1 to see if there is a learning curve.

Table 2: Session Number and Corresponding Stimuli with Noise Location and Level.

Session	Sentence #	Noise Location	SNR Level
1	1, 2	Clean-1	N/A
2	3, 4	0-1000 Hz	-20 dB
3	5, 6	1000-2000 Hz	-20 dB
4	7, 8	2000-3000 Hz	-20 dB
5	9, 10	3000-4000 Hz	-20 dB
6	1, 2	Clean-2	N/A

4. EXPERIMENTAL RESULTS

Figure 2 shows the results for Sessions 1-6. Remember that there are 25 voices and for every session, each voice is presented twice. Thus, a 90% correct means that out of 50 voice presentations, the listener identified the correct speaker 45 times. It was always a forced choice decision. Note that in Figure 2 the two lines follow the same general trend except the hearing deficit group seemed to have a larger, broader dip. Not surprising, all speakers do the best with no additive noise. Any additive noise causes a drop in performance for both groups. For the NH Group, the 2000-3000 Hz LTASS noise resulted in the lowest recognition scores. While it is tempting to say this results shows that important speaker information is in the frequency range, it is critical to complete statistical analysis both between the groups and within each group.

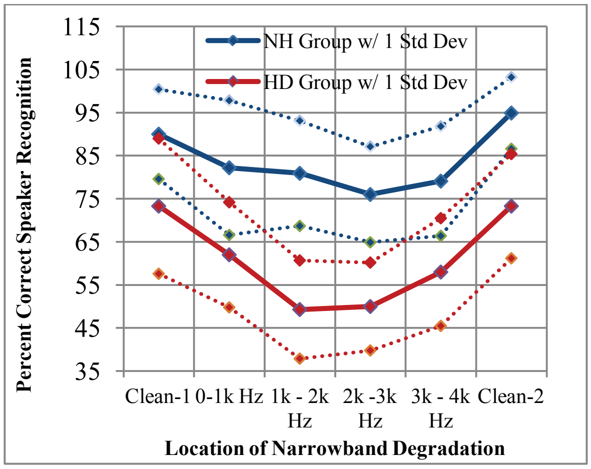

Figure 2: Correct Speaker Recognition for Sessions 1-6 with 1 Standard Deviation.

As stated before, the goal in using different sentences for each listening session was to prevent a learning curve. In hindsight, using different sentences with different noises make it difficult to compare between sessions since there is different phonetic information. However, a multi-step approach can be used to test for significant differences between the two groups and between sessions within a group.

4.1. Statistical Difference Between the NH and HD Groups

A Jarque-Bera test was used to test if each distribution is a normal distribution or not with a significance of 0.05. The null hypothesis of the Jarque-Bera test is that the distribution is a normal distribution with unknown mean and variance.

For the first results in Figure 2 (Session 1, NH = 90.0%, HD = 73.3%), the Jarque-Bera test for the NH distribution is $H_0 = 1$, which means it can be rejected at the 5% significance level. In other words, the chances of this distribution being a normal distribution is less than 5%. For the HD distribution, the Jarque-Bera test yields $H_0 = 0$ which means it cannot be rejected at the 5% significance level.

Although the Jarque-Bera test indicated that one distribution can be rejected as being normal and the other cannot, the rank sum test can be used to further validate that these distributions are statistically different. The null hypothesis of the rank sum test is that the two distributions are from identical continuous distributions with equal medians. Using the rank sum test for these two distributions with a significance of 0.05, the alternative hypothesis H_1 held true. Using the results of the Jarque-Bera test and the rank sum test, the conclusion is that the two distributions of the NH and HD for the Session 1 condition are statistically different.

For Sessions 2, 3, 4, and 6, the statistical tests yield the same results as Session 1. That is the distribution of the NH is statistically different than the HD because the two distributions fail both the Jarque-Bera test and the rank sum test. For Sessions 5 (3-4k Hz LTASS noise), both distributions passed the Jarque-Bera test and the F-Test, but then failed the T-Test (i.e., reject the null hypothesis of equal means at the 5% significance level). To summarize, the distribution of the NH and HD are statistically different for each listening condition.

The ages of NH listeners ranged from 22-49 with an average age of 34. The ages of HD listeners ranged from 49-72 with an average age of 60. For the HD, it was noted that they had failed at least one frequency presumably due to sensorineural hearing loss. In [7], the authors look at duration discrimination between young listeners and elderly listeners (both with normal hearing and hearing impairment) and concluded that age played a major role (and not hearing loss) in the diminished duration discrimination in the elderly. For these experiments, it’s hard to say if age, hearing loss, or a combination of factors contributed to the lower scores in Figure 2 for the HD Group.

4.2. Statistical Difference within the NH and HD Group

The previous section examined the statistical significance between the two listening groups for each listening scenario. This section examines the statistical significance between two listening scenario, but within the same listening group. For example, is the distribution of the 0-1k Hz LTASS noise statistical different than the 1k-2k Hz LTASS noise for the NH Group? Using the same statistical approach and the various listening permutations (i.e., clean compare to 0-1k Hz noise or, 2k-3k Hz noise compared to 3k-4k Hz noise), the only distributions for the NH Group that are statistically different with a significance of 0.05 are those listed in Table 3. This process is repeated for the HD Group and the results are also listed in Table 3. For the NH Group, all of the conditions that are statistically different involve one of the clean (no noise) conditions. Likewise, for the HD Group, most of the conditions that were statistically different involved one the clean (no noise) conditions. Any pair of conditions that are not listed in Table 3 were not statistically different for either group.

Table 3: Within Group Conditions that are Statistically Different

CONDITION 1	CONDITION 2	NH	HD
Clean-1	1k-2k Noise	X	X
Clean-1	2k-3k Noise	X	X
Clean-1	3k-4k Noise	X	
Clean-2	0-1k Noise	X	
Clean-2	1k-2k Noise	X	X
Clean-2	2k-3k Noise	X	X
Clean-2	3k-4k Noise	X	
0-1k Noise	2k-3k Noise		X

4.3. Elapsed Time

Another statistic to examine is the average time required to complete each session. In Figure 3, the average session duration time (in seconds) is plotted versus the session type. The average session duration is the time required to complete an entire listening session (50 forced choice decisions), but the duration time of the stimuli has been subtracted out. One thing to note is that for both listening groups, the average elapsed time always increased with additive noise.

Note that the two clean sessions were identical to see if there was a learning benefit. For the HD Group, the performance of correctly identifying the voices essentially had not changed between the two clean sessions. But, this graph shows that they completed the entire listening session about 73 seconds faster. Thus, they came to their conclusion much more quickly the second time for the clean session, but at a similar accuracy. For the normal listeners, they too made their decisions faster (an average of 102 seconds faster) and their performance improved by 4.9%. This increase may be due to becoming more familiar with the task, more accustomed with the GUI, and more proficient at identifying the voices.

The average session duration between the NH and HD Groups is statistically significant for each listening condition. However, within the HD Group, there is not a statistical difference between any of the listening condition. Within the NH Group, the clean-2 condition was statistically significant from the other five listening conditions.

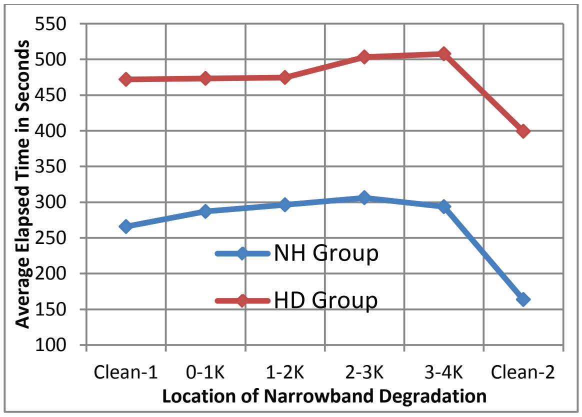

Figure 3: Average Elapsed Time vs Session Type

5. SUMMARY

The main conclusion of this research to date is that, the distribution of the NH and HD are statistically different for each listening condition, both for the performance values in Figure 2 and the elapsed time of Figure 3. Additional analysis is looking at various factors that may impact a listener's ability to identify a person's identity. Factors such as duration, amount of voiced speech, harmonic-noise-ratio, jitter, shimmer, formant locations, etc are currently being examined. Perhaps, discovering what makes a person's voice unique would enable speaker specific features.

The original goal of this effort was to discover which frequency bands are most important for the familiar speaker recognition task. As discussed in Section 5.2, all the bandlimited noise conditions resulted in lower performance compared to the clean (no noise) scenario. Yet, there was not a statistical difference between any two bandlimited LTASS noise conditions (except the last row in Table 3 for the HD Group).

While there is some research in the literature that looked at how well listeners could identify familiar speakers, the authors did not find research that looked at what frequency information was important for speaker identification. This research was a cursory look and requires more listening experiments with better randomization of stimuli and phonetic consideration.

The pure tone test was designed only to classify listeners as having normal hearing or not. It does not give insight into how good or how poor their hearing is. A listening threshold test is planned to measure this information and to see if there is a correlation between the sensorineural hearing abilities and the ability to identify familiar speakers. Additionally, a pure tone test does not take into account the cognitive reasoning process which includes such things as memory, decision making, language understanding, etc.

6. REFERENCES

[1] Crystal, A., Schimdt-Nielsen T., "Speaker Recognition by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data," *Digital Signal Processing*, 2000, Vol. 10, pp. 249-266.

[2] Calandruccio, L., Doherty, K., "Spectral Weighting for Strategies for Sentences Measured by a Correlational Method," *Journal of the Acoustical Society of America*, June 2007, Vol. 121, No. 6, pp. 2827-3836.

[3] Byrne, D., et. al., "An International Comparison of Long-Term Average Speech Spectra," *Journal of the Acoustical Society of America*, 1994, Vol. 96, No. 4, pp. 2108-2120.

[4] Coleman, R., "Speaker Identification in the Absence of Inter-Subject Differences in Glottal Source Characteristics," *Journal of the Acoustical Society of America*, 1973, Vol. 53, No. 6, pp. 1741-1743.

[5] Bricker, P., Pruzansky, S., "Effects of stimulus content on Duration on Talker Identification," *Journal of the Acoustical Society of America*, 1966, Vol. 40, pp. 1441-1449.

[6] Schmidt-Nielsen, T., Stern, F., "Identification of Known Voices as a Function of Familiarity and Narrow-Band Coding,", *Journal of the Acoustical Society of America*, Feb 1985, Vol. 77, No. 2, pp. 658-663.

[7] Van Lancker, D., Kreiman, J., "Voice Discrimination and Recognition are Separate Abilities," *Neuropsychologia,I* 1987, Vol. 25, No. 5, pp. 829-834.

[8] Lawson, A., et. al., "The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design, and Initial Findings, Interspeech 2009, pp. 1811-1814.